remote sensing

MDPI

Article

SiameseNet Based Fine-Grained Semantic Change Detection for High Resolution Remote Sensing Images

Lili Zhang [1], Mengqi Xu [2], Gaoxu Wang [3,*], Rui Shi [3], Yi Xu [3] and Ruijie Yan [2]

[1] College of Information Science and Engineering, Hohai University, Nanjing 211100, China
[2] College of Computer and Software, Hohai University, Nanjing 211100, China
[3] State Key Laboratory of Hydrology–Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China
* Correspondence: gxwang@nhri.cn

Abstract: Change detection in high resolution (HR) remote sensing images faces more challenges than in low resolution images because of the variations of land features, which prompts this research on faster and more accurate change detection methods. We propose a pixel-level semantic change detection method to solve the fine-grained semantic change detection for HR remote sensing image pairs, which takes one lightweight semantic segmentation network (LightNet), using the parameter-sharing SiameseNet, as the architecture to carry out pixel-level semantic segmentations for the dual-temporal image pairs and achieve pixel-level change detection based directly on semantic comparison. LightNet consists of four long–short branches, each including lightweight dilated residual blocks and an information enhancement module. The feature information is transmitted, fused, and enhanced among the four branches, where two large-scale feature maps are fused and then enhanced via the channel information enhancement module. The two small-scale feature maps are fused and then enhanced via a spatial information enhancement module, and the four upsampling feature maps are finally concatenated to form the input of the Softmax. We used high resolution remote sensing images of Lake Erhai in Yunnan Province in China, collected by GF-2, to make one dataset with a fine-grained semantic label and a dual-temporal image-pair label to train our model, and the experiments demonstrate the superiority of our method and the accuracy of LightNet; the pixel-level semantic change detection methods are up to 89% and 86%, respectively.

Keywords: change detection; dual-temporal remote sensing images; information enhancement; Siamese network

Citation: Zhang, L.; Xu, M.; Wang, G.; Shi, R.; Xu, Y.; Yan, R. SiameseNet Based Fine-Grained Semantic Change Detection for High Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5631. https://doi.org/10.3390/rs15245631

Academic Editors: Farid Melgani and Silvia Liberata Ullo

Received: 11 October 2023
Revised: 1 December 2023
Accepted: 1 December 2023
Published: 5 December 2023

1. Introduction

Research on change detection of HR (high spatial resolution) remote sensing images is a cross-disciplinary field that involves remote sensing technology, image processing, machine learning, deep learning, and other knowledge domains. Generally speaking, the process of extracting changed regions from two or more remote sensing images for the same location captured at different times is referred to as change detection. This technology has wide-ranging applications in land cover [1], disaster assessment [2], city management [3], ecological conservation [4], and other fields. In many countries, water shortages are becoming worse, so the monitoring of water resources and the surroundings of rivers and lakes is key for management. It is possible to monitor the construction and demolition of buildings surrounding the river or lake in a timely fashion, find illegal constructions, and prevent the illegal occupation of land resources by applying the technology of remote sensing image change detection. Hence, change detection based on remote sensing is becoming a better method to monitor changes in the surrounding rivers and lakes.

Traditional change detection is essentially a binary classification task, where each pixel in remote sensing images within the same area is classified into two categories: 'changed' and 'unchanged'. Semantic change detection attempts to further identify the

type of change that has occurred at each location. With the development of deep learning, convolutional neural networks (CNNs) have shown significant advantages over traditional methods in image processing. CNNs possess powerful feature extraction capabilities and can learn feature vectors from massive data. They can perform feature extraction and classification tasks simultaneously. Due to their impressive performance, CNNs have been widely applied in various image processing domains, including image classification, semantic segmentation, object detection, object tracking, and image restoration [5]. With the development of CNNs, change detection methods based on CNNs were proposed.

The semantic segmentation for remote sensing images aims to classify each pixel in the image to achieve image region representation. Deep-learning change detection methods based on semantic segmentation can be divided into direct comparison methods and classification-based post-processing methods [6,7]. Direct comparison methods enable real-time, end-to-end detection but are susceptible to registration accuracy and noise; in addition, they just focus on where changes happened. Classification-based post-processing methods do not require change detection labels during training and can detect pixel-level semantic changes in the images.

However, the accuracy of these kinds of change detection methods depends on the accuracy of semantic segmentation. According to the remote sensing images, there exist intra-class differences due to the complex background, different colors, and diverse shapes of the same objects, as well as inter-class similarities due to the same shapes and colors of different objects. This makes semantic change detection in remote sensing images challenging. Therefore, we explore a lightweight semantic segmentation network (LightNet) to carry out the pixel-level semantic classification, use the parameter-sharing SiameseNet as the network architecture to get the different classifications of each pixel in the image pair, and finish with pixel-level change detection based on semantic comparison of the image pair.

The main contributions of our work are as follows:

(1) We propose a lightweight parameter-sharing SiameseNet to solve the semantic classification of each pixel in the image pair, formalizing the pixel-level semantic comparison into a set operation problem to carry out pixel-level semantic change detection of the image pair directly.

(2) We propose a lightweight semantic segmentation network (LightNet), which consists of four long–short branches, each of which includes lightweight dilated residual blocks and a channel or spatial information enhancement module.

(3) The feature information is transmitted, fused, and enhanced simultaneously among the four branches. Two branches carry out the fusion and channel information enhancement of two large-scale feature maps, while the other two carry out the fusion and spatial information enhancement of two small-scale feature maps. The four feature maps are concatenated to form the input of the Softmax.

(4) We made a fine-grained dataset for the training of SiameseNet. Each sample in the dataset has two annotations, a semantic label and a matching label for dual-temporal images, which guarantee the input of SiameseNet to be dual-temporal image pairs with the semantic label.

2. Related Work

2.1. Change Detection of Remote Sensing Images

The existing change detection methods can be divided into the image difference method [8,9], change vector analysis (CVA) [10,11], principal component analysis (PCA) [12,13], and the deep learning method [6,7,14–27].

The image difference method refers to subtracting the bands of dual-temporal images to obtain the difference map. This method is very simple and divides the image pixels into two results: change or not change. Change vector analysis (CVA) is an extension of the image difference method. It uses the information of multiple bands to obtain the change vector with length and direction. The length of the vector represents the intensity of the change, and the direction of the vector represents the change type. Principal component

analysis (PCA), also known as eigenvector transformation, is a technology used to reduce the dimension of datasets. These change detection methods have low detection accuracy and the boundary between the detected changed region and the unchanged is rough.

Recently, deep learning has developed rapidly, and many remote sensing image change detection methods based on CNNs came into being. The remote sensing image change detection method based on deep learning directly learns the change features from the dual-temporal images, segments the image through the change features, and finally obtains the change map. Zhang et al. [14] proposed a feature difference convolutional neural network-based change detection method that achieves better performance than other classic approaches and has fewer missed detections. Daudt et al. [15] proposed different methods, named as FC-EF, FC-Siam-conc, and FC-Siam-diff, sequentially referring to U-Net, which verified the feasibility of a fully convolutional network for change detection. Chen et al. [16] proposed STANet, which establishes the spatiotemporal relationship between multitemporal images by adding two spatiotemporal attention modules. Experiments show that the attention module of STANet can reduce the detection error caused by improper registration of multitemporal remote sensing images. Ke et al. [17] proposed a multi-level change context refinement network (MCCRNet), which introduces a change context module (CCR) to capture denser change information between dual-temporal remote sensing images. Peng et al. [18] proposed a difference-enhancement dense-attention convolutional neural network (DDCNN), which combines dense attention and image difference to improve the effectiveness of the network and its accuracy in extracting the change features.

However, the above change detection methods actually complete a binary classification task. Each pixel on the remote sensing image is classified into 'changed' and 'unchanged', which does not identify the semantic information of the change parts.

In order to obtain the change region and its semantic information, semantic change detection has gradually come to people's attention. Semantic change detection can be categorized into three types: prior-based semantic change detection [19], multitask model-based semantic change detection [6,20], and semantic segmentation-based semantic change detection [7,21]. Prior-based methods require the collection of prior information. A prior semantic information-guided change detection method, PSI-CD, was introduced by [19], incorporating prior information for semantic change detection. This approach effectively mitigates the model's dependence on datasets, creating semantic change labels on public datasets and achieving semantic change detection in dual-temporal high resolution remote sensing images. Multitask models handle semantic segmentation and change detection in parallel. Daudt et al. [6] proposed integrating semantic segmentation and change detection into a multitask learning model, and the association between the two subtasks is also considered in the model to some extent. A dual-task semantic change detection network (GCF-SCD-Net), introduced by [20] utilizes a generated change field (GCF) module for the localization and segmentation of changed regions. Semantic segmentation-based approaches do not explicitly emphasize simultaneous handling, which can be categorized into direct comparison methods and classification-based post-processing methods [7,21].

However, these methods commonly ignore the inherent relationship between the two subtasks and encounter challenges in effectively acquiring temporal features [7,22,23]. To solve this problem, a semantic change detection model based on the Siamese network has emerged on the basis of semantic segmentation based change detection, which uses Siamese networks to extract dual-temporal image features [22,23].

Yang et al. [24] found that the change of land cover appears in different proportions in multitemporal remote sensing images and proposed an asymmetric Siamese network to complete semantic change detection. Peng et al. [25] proposed SCDNet, which realizes end-to-end pixel-level semantic change detection based on Siamese network architecture. Fang et al. [26] proposed a densely connected Siamese network (SNUNet-CD), which decreases the uncertainty of the pixels at the edge of the changed target and the determination miss of small targets. Chen et al. [27] proposed a bitemporal image transformer (BIT) and incorporated it in a deep feature differencing-based CD framework. This method used only

three-fold lower computational costs and model parameters, significantly outperforming the purely convolutional baseline.

Now, semantic change detection is being applied in water resource management in China, but the accuracy and testing time of existing methods should be improved, so we propose a pixel-level fine-grained change detection method for remote sensing images to meet the needs of monitoring water resources.

2.2. Semantic Segmentation

Semantic segmentation aims to determine the label of each pixel in the image, so as to realize the region division of the image. Early image semantic segmentation methods mainly use manual extraction of some shallow features, such as edge [28], threshold [29], etc. However, for complex scene images, the expected effect of segmentation cannot be achieved. In recent years, the semantic segmentation method based on deep learning has achieved outstanding performance.

Long et al. [30] proposed FCN, which extends the new idea of deep learning in the field of image segmentation and realizes end-to-end pixel-level semantic segmentation. Noh et al. [31] proposed Deconvnet, which adopts a symmetrical encoder–decoder structure to optimize the FCN. Badrinarayanan et al. [32] proposed SegNet, which carries out maximum unpooling in the decoding part to realize upsampling; this improves the segmentation accuracy compared with FCN. Zhang et al. [33] proposed an FCN without pooling layers, which can achieve higher accuracy in extracting tidal flat water bodies from remote sensing images. U-Net, proposed by Ronneberger et al. [34], can train pictures in the form of end-to-end when there are few pictures in the dataset. Li et al. [35] proposed a Multi-Attention-Network (MANet), which optimizes the U-Net by extracting contextual dependencies through multiple efficient attention modules. Ding et al. [36] proposed a local attention network (LANet), which improves semantic segmentation by enhancing feature representation by integrating a patch attention module and an attention embedding module into a baseline FCN. Zhang et al. [37] proposed a multiscale contextual information enhancement network (MCIE-Net) for crack segmentation and redesigned the connection structure between the U-Net encoder and decoder to capture multiscale feature information, enhancing the decoder's fine-grained restoration ability of crack spatial structure. He et al. [38] proposed Mask R-CNN, a network model combining target detection and semantic segmentation, so the model can classify, recognize and segment images. The DeepLabv3+ network proposed by Chen et al. Zhang et al. [39] improved Mask R-CNN for high spatial resolution remote sensing images building extraction. The latest and best network framework of the DeepLab series [40] is based on an encoder–decoder structure and atrous spatial pyramid pooling (ASPP). It achieved a brilliant performance on the PASCAL-VOC2012 dataset. Different from the current popular serially connected network, the HRNet proposed by KeSun and others [41] is a new parallel architecture. It continuously fuses with each other in four stages to maintain the resolution in the whole process and avoid the loss of information caused by downsampling. Therefore, the predicted image is more accurate in space. However, the complex parallel subnet and repeated feature fusion of HRNet lead to a huge number of parameters and computational complexity. In high resolution remote sensing images, the intra-class differences are significant due to complex scenes, large-scale changes, different colors, and diverse shapes. On the other hand, different classes exhibit similarities in terms of shapes and colors, resulting in small inter-class differences [42]. These factors pose significant challenges for semantic segmentation in high resolution remote sensing imagery and lead to low recognition accuracy of existing semantic segmentation models.

3. Dataset

The public dataset for semantic change detection is absent, and the semantic label of the public dataset HRSCD is rough and not applied to the refined management of government. The sample in HRSCD is illustrated in Figure 1, which shows that the change label is too coarse to be trained to carry out the pixel-level change detection. Hence, we

make a fine-grained dataset for the training of SiameseNet based on semantic segmentation, and each sample in the dataset has two annotations, semantic label and image pair label, which guarantee the input of HR remote sensing image pairs and the semantic segmentation of the image pair simultaneously.

(**a**) (**b**) (**c**) (**d**)

Figure 1. A sample in HRSCD. (**a**) Semantic label of pre-temporal image. (**b**) Semantic label of post-temporal image. (**c**) Pre-temporal image. (**d**) Post-temporal image. (Artificial surfaces: Agricultural areas: Forests:).

In order to demonstrate the accuracy of any change detection method based on SiameseNet, pixel-level precise matching of the image pair is necessary. And different semantic change detection tasks generally need different semantic labels in application, so we propose an efficient pixel-level precise matching algorithm to solve the matching problem of large-scale remote images in applications quickly and accurately.

3.1. Fast and Precise Matching Algorithm for Large-Scale Remote Sensing Dual-Temporal Images

The remote sensing images from different satellites appear at varying scales, such as an average size of 30,000 × 30,000 and spatial resolution of 4 m for GF-2, and an average size of 7800 × 7800 and spatial resolution of 30 m for Landsat, but they all have latitude and longitude label information. Therefore, we combine latitude and longitude information for large size matching and pixel information for point matching to achieve fast and precise matching for large-scale remote sensing image pairs [43,44]. The method is as follows.

(1) We firstly take one original remote sensing image pair from a satellite, and extract the time label and the latitude and longitude coordinates of the four corner points, which are denoted clockwise: a_{m1k1}, a_{m2k2}, a_{m3k3}, a_{m4k4}, and b_{p1q1}, b_{p2q2}, b_{p3q3}, b_{p4q4}, respectively, corresponding to the pre-temporal and the post-temporal remote sensing image. Secondly, we calculate the intersection of the pixel regions A enclosed by a_{m1k1}, a_{m2k2}, a_{m3k3}, a_{m4k4} and the pixel region B enclosed by b_{p1q1}, b_{p2q2}, b_{p3q3}, b_{p4q4}, and obtain the latitude and longitude coordinates of the four intersection points: c_{m1k1}, c_{m2k2}, c_{m3k3}, c_{m4k4}. Subsequently, the pixels outside the region enclosed by c_{m1k1}, c_{m2k2}, c_{m3k3}, c_{m4k4} are removed from the original remote sensing image pair and we then sort the remaining pixels into a raster format image. Based on latitude and longitude calculations, this approach achieves fast large-scale matching of remote sensing image pairs.

(2) We refine the matching based on a SIFT+ matching algorithm that we propose. We use the Scale-Invariant Feature Transform (SIFT) to obtain the matched feature points and denote them as S, and we set a distance threshold T for the matched point pairs to refine the matching result.

(3) Based on the fundamental theory that any three randomly selected points from the sample cannot be collinear, at least four sample data points are randomly selected from the set of matched feature points S. These selected matching points are used as an initial set to calculate the distance matrix between the corresponding matched feature points.

(4) The remaining matched points in set S are used to calculate new position coordinates based on the average distance in the distance matrix. The distance d between the calculated position coordinates and the original position coordinates is then computed.

(5) If d $\geq T$, the matched point is defined as an incorrect match. If d < T, the matched point is defined as a correct match.

(6) Repeat steps (3) to (5) until the root mean square error of the matched point pairs meets our requirement. Select the group with the highest number of correct matches as the final set of correct matching points, and we achieve fast and precise pixel-level matching of the large-scale remote sensing image pairs.

The comparison of our method, the SIFT+ algorithm, and the SIFT algorithm is shown in Table 1. It is seen from Table 1 that our method is more accurate.

Table 1. Comparison of our method SIFT+ algorithm and SIFT algorithm.

Algorithm	Number of Matching Point Pairs	MSE
SIFT	75	2.395
SIFT+	34	0.842

3.2. Data Making with Semantic and Matching Labels

(1) We firstly finish the matching labels for the dual-temporal remote sensing images based on the matching result of the method in Section 3.1, and we use i and i+ to demote them respectively, i = 1,2,. . ., where i corresponds to the pre-temporal image and i+ corresponds to the post-temporal image. These symbol pairs can express matching and timing relationships of the image pair simultaneously.

(2) We make the semantic labels semi-automatically for each matched image. We find enough points through the edge of any object manually, and then use linear interpolation to draw an enclosed area and define the same semantic labels for the enclosed interior automatically. Our method makes data faster than those manual methods.

(3) We obtain a new dataset with semantic and matching labels for change detection of the remote sensing images, and it is shown that the making process of our dataset can be applied to any semantic change detection task based on deep learning. Moreover, the samples in our dataset have better fine-grained semantic information than the public dataset HRSCD, and we can achieve fine-grained semantic change detection for remote sensing images based on the dataset we made.

We construct the dataset using two Gaofen-2 remote sensing images taken in Dali, Yunnan Province, on 13 February 2017 and 1 April 2020, respectively. Due to the disorderly tourism development of Lake Erhai and lax law enforcement on illegal buildings in recent years, more and more buildings have been constructed surrounding Erhai Lake, leading to poor water quality. If the management finds illegal occupation based on remote sensing images in time, it will be helpful to protect the environment and so on. Therefore, we select some key changes in water management as a case study of our method in this paper. We obtain 2100 pairs of change detection samples with semantic and matching labels. The example in our dataset is shown in Figure 2. It is seen that our data has better fine-grained semantic information than the data in the public dataset HRSCD.

(**a**) (**b**) (**c**) (**d**)

Figure 2. The sample in our fine-grained semantic change detection dataset. (**a**) Semantic label of pre-temporal image. (**b**) Semantic label of post-temporal image. (**c**) Pre-temporal image. (**d**) Post-temporal image. (Building: Plant: Bare soil: Water: Background:).

4. Methodology

4.1. Framework

Our framework is a deep learning model based on the Siamese network for the semantic change detection of multi-source remote sensing images. The key to the Siamese network is two parallel semantic segmentation networks, which obtain the different semantic classifications of the image pair via parameter-sharing and finish with the pixel-level change detection based on semantic comparison of the image pair. The framework of our method is shown in Figure 3.

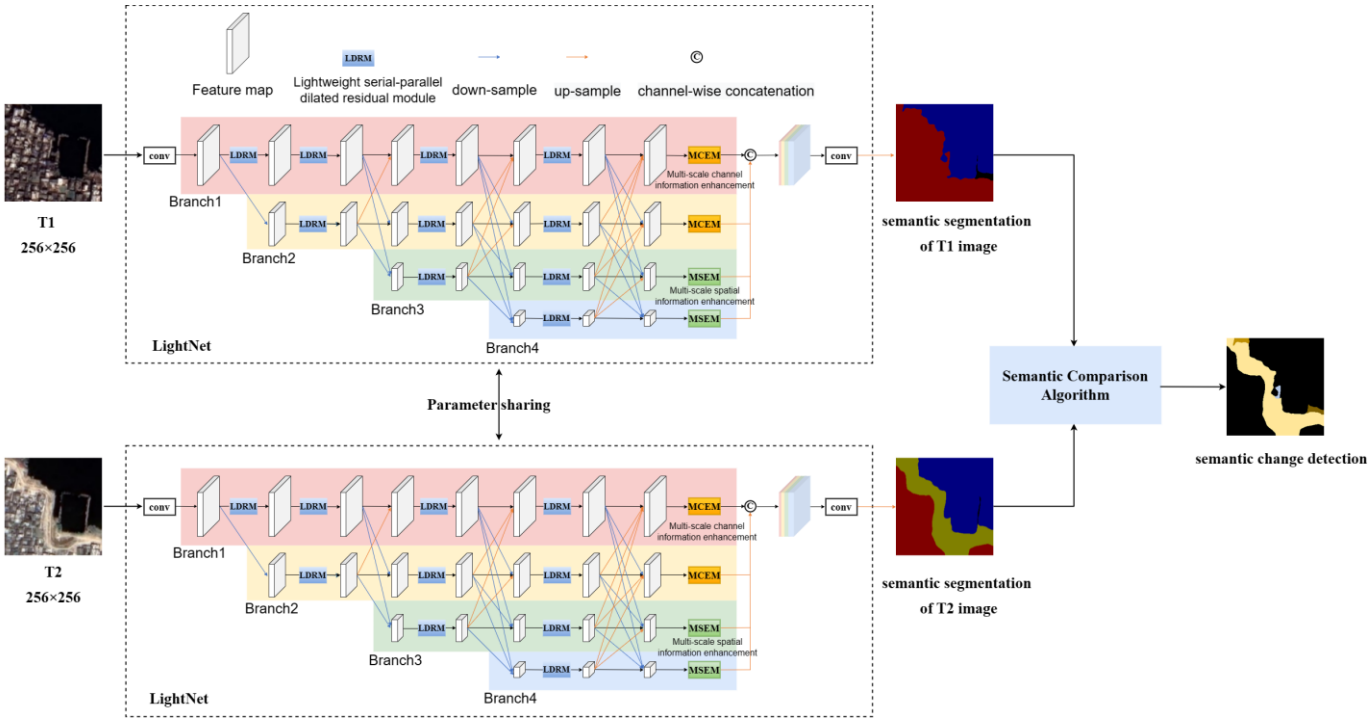

Figure 3. The framework of our method.

4.2. LightNet

The semantic model is the key to improving the accuracy of our method. We name our network LightNet because it is an efficient lightweight semantic segmentation network. The LightNet consists of four long–short branches. Two branches include lightweight serial-parallel dilated residual modules (LDRM) and a multiscale channel information enhancement module (MCEM), while the other two branches include LDRM and a multiscale spatial information enhancement module (MSEM). The LDRM integrates the advantages of serial and parallel dilated residual networks and has a lightweight serial-parallel structure. The MCEM captures the correlation between local and global features by calculating multiscale spatial attention matrices and performs weighted fusion of the upsampled multiscale spatial features extracted by the LDRM to enhance the semantic consistency of discriminative features on the same object. The MSEM utilizes attention mechanisms to compute channel weight vectors and performs weighted fusion of the same-channel features at different scales to enhance the semantic distinctiveness among different objects.

The LightNet structure is shown in Figure 3. The first branch of the backbone network consists of four LDRMs and an MCEM, the second branch includes three LDRMs and an MCEM, the third branch includes two LDRMs and an MSEM, and the fourth branch includes an LDRM and an MSEM. Different feature maps from four branches are upsampled or downsampled and then input into other branches and take part in the fusing operation.

4.2.1. Lightweight Serial-Parallel Dilated Residual Module (LDRM)

In order to extract the required multiscale contextual information for the MSEMs and MCEMs, we integrate the advantages of serial and parallel dilated residual networks and design the lightweight serial-parallel dilated residual module (LDRM). The specific structure of the LDRM is shown in Figure 4. It consists of three dilated residual blocks with parameter sharing, so achieves a lightweight serial-parallel structure. The three parallel dilated residual blocks of the LDRM consist of three, two, and one dilated residual layers respectively, and they are series connections when there are at least two. Each dilated residual block is composed of two convolutional layers and a skip connection. The three blocks extract multiscale spatial features and channel features using different dilation rates. The blocks share parameters to achieve a lightweight serial-parallel structure. The multiscale spatial features, channel features, and original features in each branch are fused together as the final output of this module. We will give the specific calculation in detail.

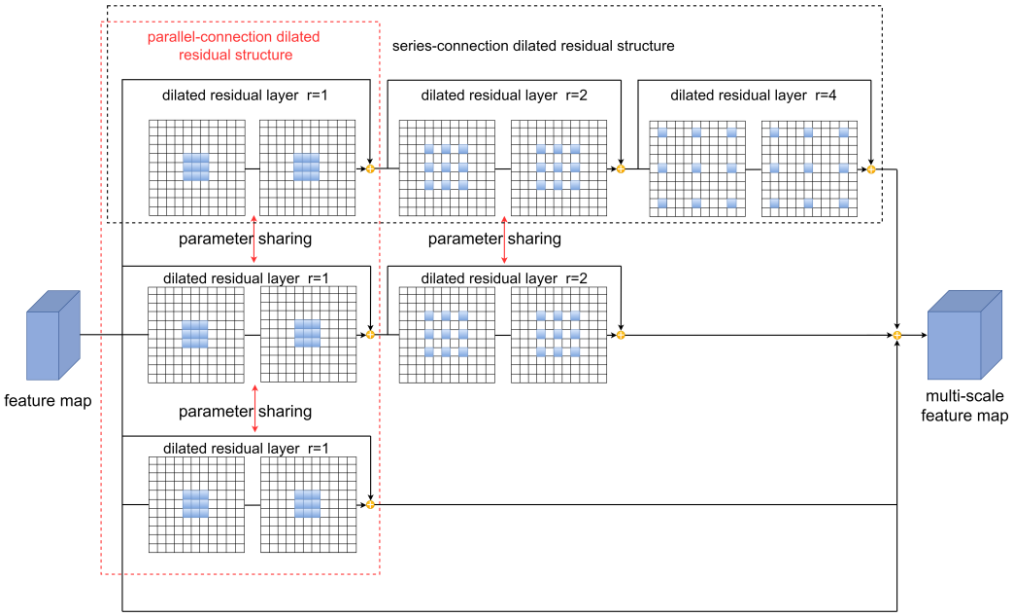

Figure 4. Lightweight serial-parallel dilated residual module.

The first block consists of three series-connection dilated residual layers, with dilation rates of 1, 2, and 4, respectively. Its output is represented as follows:

$$Output1 = D_4(D_2(D_1(X))) \tag{1}$$

where D_i represents the output of the dilated residual block with dilation rate i, and X represents the feature map as the input.

The second block consists of two series-connection dilated residual layer with dilation rate of 1 and 2, respectively. Its output is represented as follows:

$$Output2 = D_2(D_1(X)) \tag{2}$$

The third block consists of a dilated residual layer with a dilation rate of 1, and its output is represented as $D_1(X)$. The final output of the LDRM is the fusion of the multiscale features extracted from the three dilated residual blocks and the original input features, represented as:

$$Output = D_4(D_2(D_1(X))) + D_2(D_1(X)) + D_1(X) + X \tag{3}$$

The parameter-sharing among the three parallel dilated residual blocks is illustrated in Figure 4.

4.2.2. Multiscale Spatial Information Enhancement Module (MSEM)

In order to capture more correlation between local and global features and enhance the semantic consistency of discriminative features, a multiscale spatial information enhancement module (MSEM) is designed. The module is divided into two parts, with one solving the multiscale spatial attention matrix and the other performing a weighted fusion of the multiscale spatial features based on the multiscale spatial attention matrix. The MSEM structure, shown in Figure 5, improves the network performance to distinguish confused categories easily via enhancing the semantic consistency of the same object.

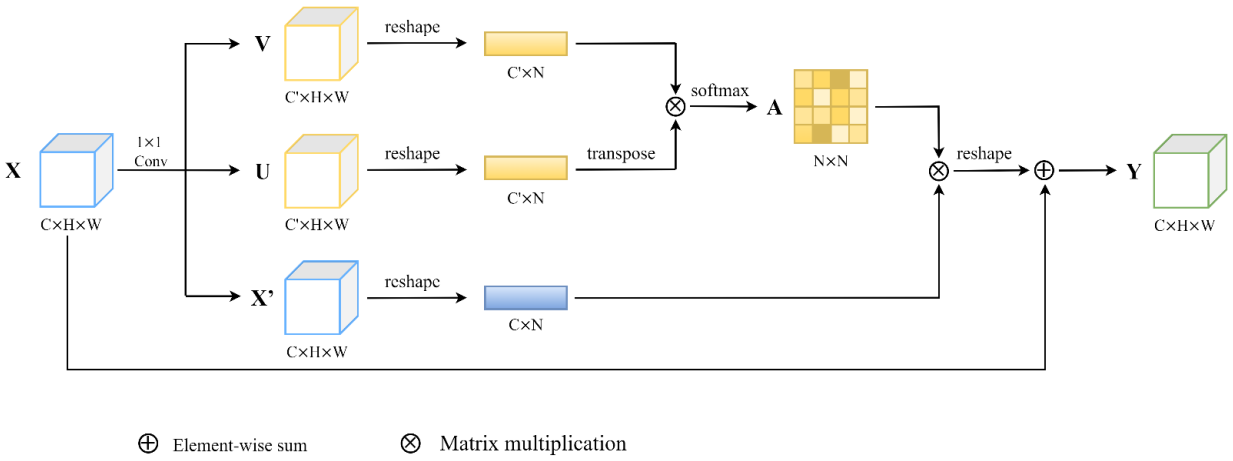

Figure 5. Multiscale spatial information enhancement module.

The computation of the multiscale spatial attention matrix is performed as follows: Firstly, the multiscale spatial feature map $X \in \mathbb{R}^{C\times H\times W}$ (where C, H, W represent the number of channels, height, and width of the feature map, respectively) is passed through a convolutional layer (C, 1×1) to obtain the feature map $X' \in \mathbb{R}^{C\times H\times W}$. Similarly, X is passed through another convolutional layer (C', 1×1) to obtain two feature maps U, $V \in \mathbb{R}^{C'\times H\times W}$, where C' is a factor of C. Secondly, the three-dimensional matrix X' is reshaped into a two-dimensional matrix $C \times N$, and the three-dimensional matrices U and V are also reshaped into two-dimensional matrices $C' \times N$, where $N = H \times W$. Next, the transpose of the matrix U is multiplied by the matrix V, and the obtained matrix is passed through the Softmax function to compute the spatial attention matrix $A \in \mathbb{R}^{N\times N}$. The calculation formula is as follows:

$$A = softmax\left(U^T V\right) \tag{4}$$

The calculation of the weighted fusion of multiscale spatial features based on the multiscale spatial attention matrix is as follows: The two-dimensional matrix X' is multiplied by the transpose of the spatial attention matrix A, and the obtained two-dimensional matrix is reshaped into a three-dimensional matrix to obtain the sum of it and the input feature map X by element-wise sum, so we obtain the final enhanced feature map $Y \in \mathbb{R}^{C\times H\times W}$. The calculation formula is as follows:

$$Y = r\left(X' A^T\right) \bigoplus X \tag{5}$$

where r represents the reshape operation, and $\bigoplus$ represents element-wise sum.

4.2.3. Multiscale Channel Information Enhancement Module (MCEM)

In order to enhance the semantic differences among different objects and alleviate the information interference caused by the similarity of different object classes, a multiscale channel information enhancement module (MCEM) was designed, which consists of two parts: One part is to compute the weight vectors of different channels using attention

mechanisms, and the other part is for the weighted fusion of the different scales' features in the same channel. Its structure is shown in Figure 6.

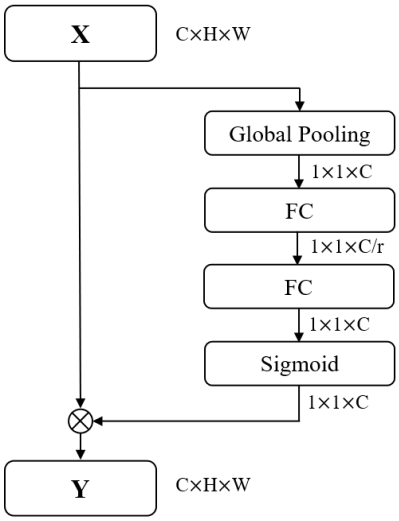

Figure 6. Multiscale channel information enhancement module.

The computing operation of the weight vectors of all the channels is as follows: Firstly, global average pooling is applied to the input feature map $X \in \mathbb{R}^{C \times H \times W}$, and it is compressed into a global spatial feature with size $1 \times 1 \times C$. Secondly, the global spatial feature is passed through two fully connected layers and a sigmoid activation function to obtain a channel weight vector with size $1 \times 1 \times C$. Each element in the channel weight vector corresponds to the weight of a feature channel, ranging from 0 to 1. In the first fully connected layer, the number of channels is reduced to $\frac{C}{r}$, where r represents the scale factor.

According to the channel weight vector, different scale features in the same channel are weighted and fused. Specifically, each element in the channel weight vector is multiplied by the corresponding channel in the original feature map, and common channel features are ignored to enhance semantic differences and alleviate information interference caused by similarity among different objects. The weighted fusion calculation formula is as follows:

$$Y_c = z_c M_c, \ c = 1, 2, \ldots, C \tag{6}$$

where Y_c represents the feature map outputted from the cth channel, M_c represents the feature map input to the c-th channel, and z_c is the channel weight vector of the feature map in the c-th channel.

4.3. Loss Function for LightNet

Loss function is one of the most important parts in deep learning because it guides the CNNs to optimize model parameters during the back-propagation period. The loss function of LightNet we designed is represented as follows:

$$Loss = Loss1 + Loss2$$

$Loss1$ and $Loss2$ are multi-class cross-entropy loss functions that evaluate the loss between the predicted semantic segmentations of the image pair and the ground truths. They are defined as follows:

$$Loss1 = -\frac{1}{m}\sum_{j \in P1}\sum_{i=1}^{n} I\left(y^j = i\right) \times log\left(P\left(y^{(j)} = i | x^{(j)}\right)\right) \tag{7}$$

$$Loss2 = -\frac{1}{m}\sum_{j \in P2}\sum_{i=1}^{n} I\left(y^j = i\right) \times log\left(P\left(y^{(j)} = i | x^{(j)}\right)\right) \tag{8}$$

where $P1$ and $P2$ represent the pre-temporal and post-temporal remote sensing images, respectively, m denotes the number of pixels in the remote sensing images, and n represents the number of classes. $I(x)$ is an indicator function that will be 1 when the predicted class y^j of pixel j matches the true class i, and 0 otherwise. $P\left(y^{(j)} = i|x^{(j)}\right)$ represents the probability that pixel j belongs to class i, which can be obtained via a Softmax classifier.

4.4. Semantic Comparison Algorithm

Following the parallel semantic segmentation of dual-temporal remote sensing images based on LightNet, we perform the semantic comparison algorithm for the two different semantic segmentations, and then directly express the semantic changes of the land cover in one remote sensing image.

Our algorithm is as follows:

The inputs are the pixel-level semantic segmentation results of the pre-temporal and post-temporal remote sensing images, denoting the semantics in the pre-temporal image as x_i, and the matching semantics in the post-temporal image as y_i; for each matching pixel pair (x_i,y_i), if the prediction category of x_i is the same as y_i, the semantic label remains as x_i. Otherwise, the semantic category of x_i is replaced with the semantic change label $x_i \rightarrow y_i$. This method can express the pixel-level semantic change intelligently and help the management find the specific change without any manual semantic comparison. The visualization map of semantic change labels included in this paper is shown in Figure 7. We have five class objects (building, plant, bare soil, water, and background) and 21 semantic changes in our dataset.

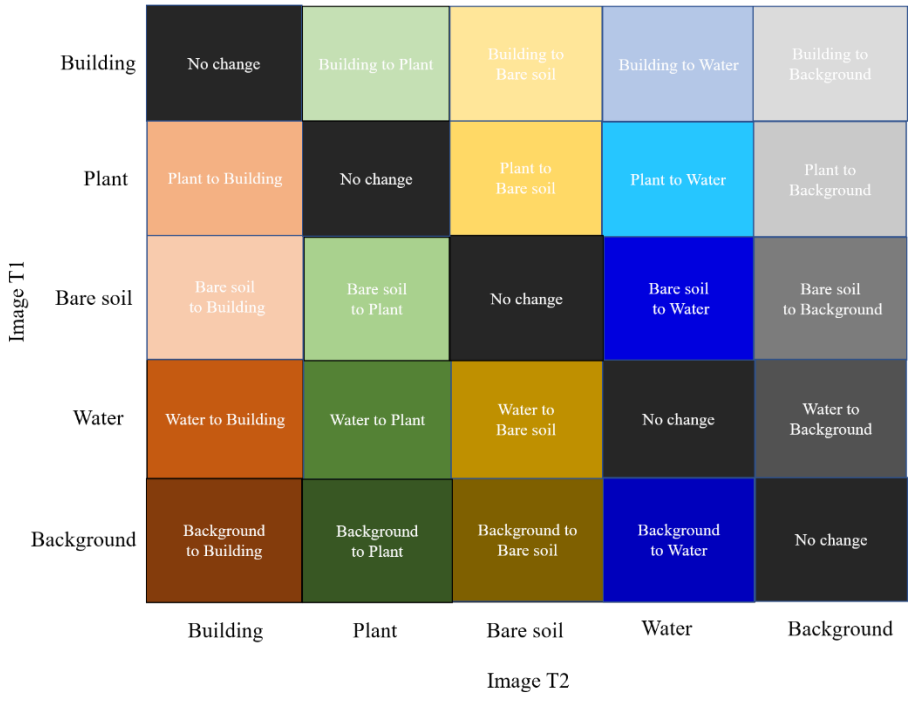

Figure 7. The visualization map of object semantic changes.

5. Experiment and Analysis

5.1. Experimental Setup

The experiments are implemented on a system with NVIDIA GeForce RTX 2060 and Intel(R) Core (TM) i7, and the operating system is Windows 10. The software environment of the system is ENVI 5.3, Python 3.8, and Pytorch 1.8.1. After testing experiments, we set the model training parameters as follows: The batch size is 4, the learning rate is 0.001, the epoch is 100, the momentum is 0.9, the weight decay is 0.0005, and the optimizer is Adam.

5.2. Evaluation Metrics

(1) Semantic segmentation metrics

The performance evaluation indexes of general semantic segmentation model mainly include mean pixel accuracy (*MPA*) and mean intersection over union (*mIoU*). In order to accurately analyze the experiments, these two indicators are selected to quantitatively evaluate the model.

$$m_i = \sum_{j=1}^{N} n_{ij} \tag{9}$$

$$MPA = \frac{1}{N} \sum_{i=1}^{N} \frac{n_{ii}}{m_i} \tag{10}$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{n_{ii}}{m_i + \sum_{j=1}^{N} n_{ji} - n_{ii}} \tag{11}$$

where N represents the total number of categories, n_{ij} represents the number of pixels which should be in class i but is classified as class j, and m_i represents the total number of pixels of class i.

(2) Change detection metrics

To evaluate the efficiency of change detection and consider whether the pixels have changed, we selected the *PA* (Pixel Accuracy) as the evaluation index.

$$PA = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

5.3. Evaluation Metrics

5.3.1. Performance Analysis of LightNet

In order to verify the effectiveness of the semantic segmentation model LightNet, we selected U-Net, PSPNet, and DeepLabv3+ for comparison experiments. The semantic segmentation results of remote sensing images are shown in Figure 8. Figure 8a shows the original image, Figure 8b shows the ground truth, and Figure 8c–f show the semantic segmentation experiments obtained by U-Net, PSPNet, DeepLabv3+, HRNet, and our model, respectively. The semantic segmentation of our model is closer to the ground truth than the other models, so our model is better. In remote sensing images including rivers and lakes, the difference of various buildings is significant, while the differences of plants, bare soil, and water bodies are not. However, our model uses LDRM to obtain more contextual information, enhances the semantic consistency of the discriminative features, and introduces MSEM and MCEM to enhance the model's ability to distinguish confused categories easily. The semantic segmentations of our model are significantly better in terms of overall and smoothness than other models, which indicates LightNet is superior to other semantic segmentation models and can effectively solve intra-class inconsistency and inter-class similarity in remote sensing images.

The quantitative comparison of the four methods for semantic segmentation on *PA*, *IoU*, *mIoU*, and *MPA* is shown in Table 2. Compared with U-Net, PSPNet, DeepLabv3+, and HRNet, our model has advantages for semantic segmentation in remote sensing images. The *mIoU* and the *IoU* of each category of LightNet for semantic segmentation is the highest. The *PA* on building, plant, bare soil, and water has been improved by at least 2.2%, 2.1%, 1.3%, and 2.1%, respectively, and the mean *PA* is improved by 1.9%. According to the network structure, the parameter number of Lightnet is 23M, and the HRNet is 29M, so our model is lighter, and the testing time of our model is about 0.027 s, which means the model can deal with 36 frames per second.

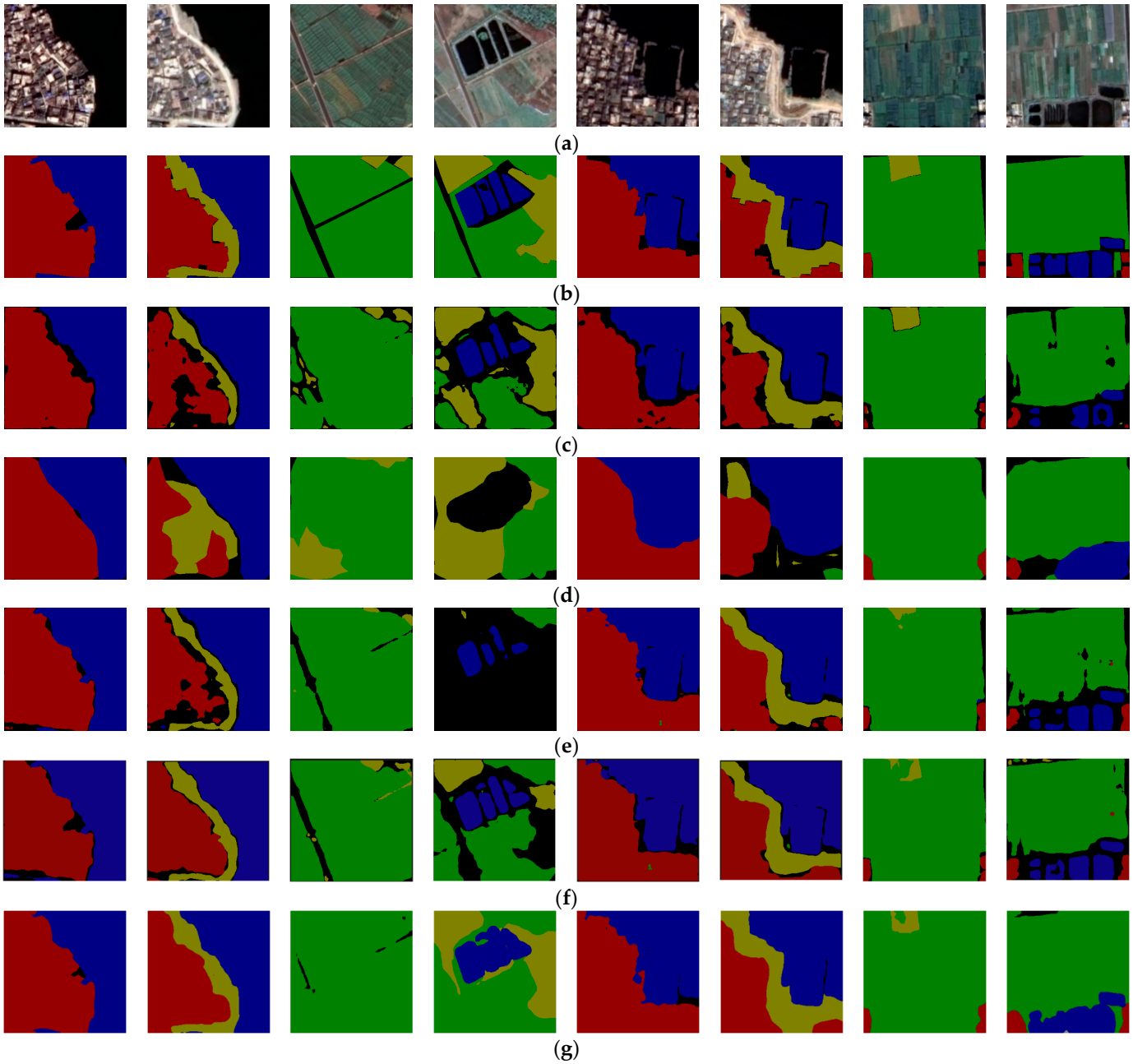

Figure 8. Experiments of different methods for semantic segmentations. (**a**) Original image. (**b**) Ground truth. (**c**) U-Net. (**d**) PSPNet. (**e**) DeepLabv3+. (**f**) HRNet. (**g**) LightNet. (Building: Plant: Bare soil: Water: Background:).

Table 2. Comparison of five methods for semantic segmentation. (%).

Method	Building		Plant		Bare Soil		Water		Average	
	IoU	*PA*	*IoU*	*PA*	*IoU*	*PA*	*IoU*	*PA*	*mIoU*	*MPA*
U-Net [34]	67.2	80.5	80.4	87.9	46.2	65.3	85.6	90.3	69.9	81.0
PSPNet	73.7	85.5	80.6	89.1	43.8	62.3	88.7	92.7	71.7	82.4
DeepLabv3+ [40]	76.0	86.2	81.2	89.3	58.7	71.5	90.3	93.2	76.6	85.1
HRNet [41]	78.2	88.5	82.4	90.4	60.2	74.3	91.2	95.3	78.0	87.1
LightNet	78.7	90.7	85.3	92.5	61.4	75.6	93.0	97.2	79.6	89.0

5.3.2. Ablation Study on LightNet

We decompose the network step by step and verify the performance of each optimized module in LightNet. The performance evaluations are shown in Table 3, in which we use the bounding boxes to highlight and demonstrate the differences of semantic segmentation with different optimization technologies in Figure 9, which easily indicate the effectiveness of different modules. Model1 represents the HRNet model without any improvement strategies. Model2 incorporates the LDRM module to capture more multiscale contextual information. Compared to Model1, Model2 improves accuracy by 1.1%. Based upon Model2, Model3 introduces MSEM for multiscale spatial information enhancement, to capture correlations between local and global features. As a result, Model3 improves segmentation accuracy by 0.5% compared to Model2. Model4, in comparison to Model3, incorporates MCEM to enhance the representation of key channel features. This helps alleviate information interference caused by similarities of different categories. Model4 achieves a segmentation accuracy of 89%. These demonstrate that each optimized module can enhance the performance of semantic segmentation in remote sensing images.

Table 3. Comparison of four methods for semantic segmentation. (%).

Model	LDRM	MSEM	MCEM	MPA
Model1				87.1%
Model2	√			88.2%
Model3	√	√		88.7%
Model4	√	√	√	89.0%

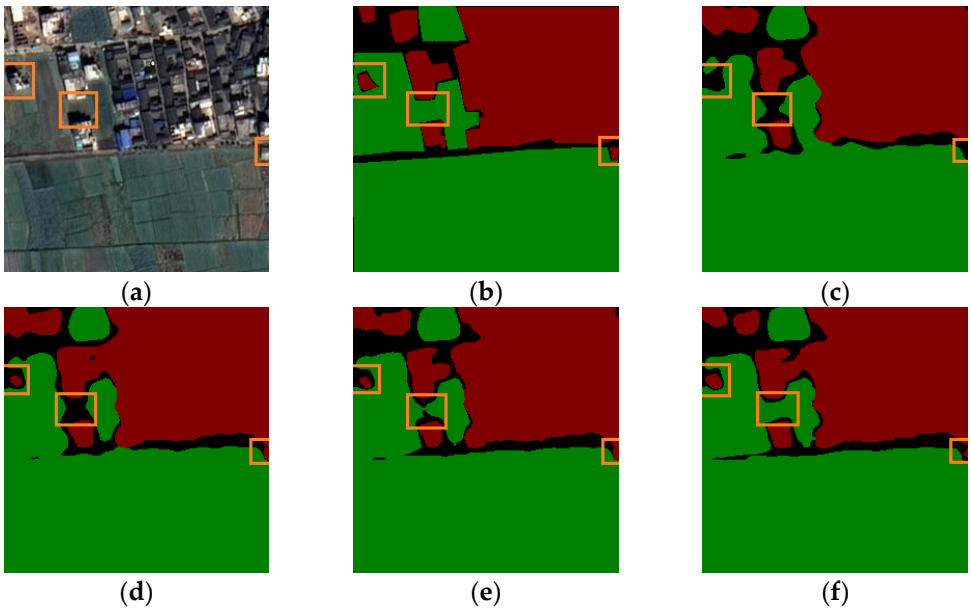

Figure 9. Ablation experiments of LightNet. (**a**) Original image. (**b**) Ground truth. (**c**) HRNet. (**d**) HRNet + LDRM. (**e**) HRNet + LDRM + MSEM. (**f**) HRNet + LDRM + MSEM + MCEM. (There are missed or false detections inside the orange box.)

5.3.3. Performance Analysis of Change Detection Method

We verified the efficiency of the semantic segmentation model LightNet we proposed. Now we verify the performance of our pixel-level semantic change detection method based on LightNet and the SiameseNet framework. The change detection results of different methods are shown in Figure 10. The change detection result obtained by our method is better and more consistent in the visual interpretation. The most changes in the first and third image pair are in building→bare soil, which indicates that there were some illegal buildings before, and the most changes in the second and fourth pair in Figure 10 are in

plant→water, which indicates that there were some illegal occupations of water resources before. Besides the fine-grained semantic change detection, Table 4 shows that the accuracy of our method for change detection is the highest in the four methods by comparing the binary accuracy of the change region, which demonstrates the superiority of our method in fine-grained semantic change detection in remote sensing images.

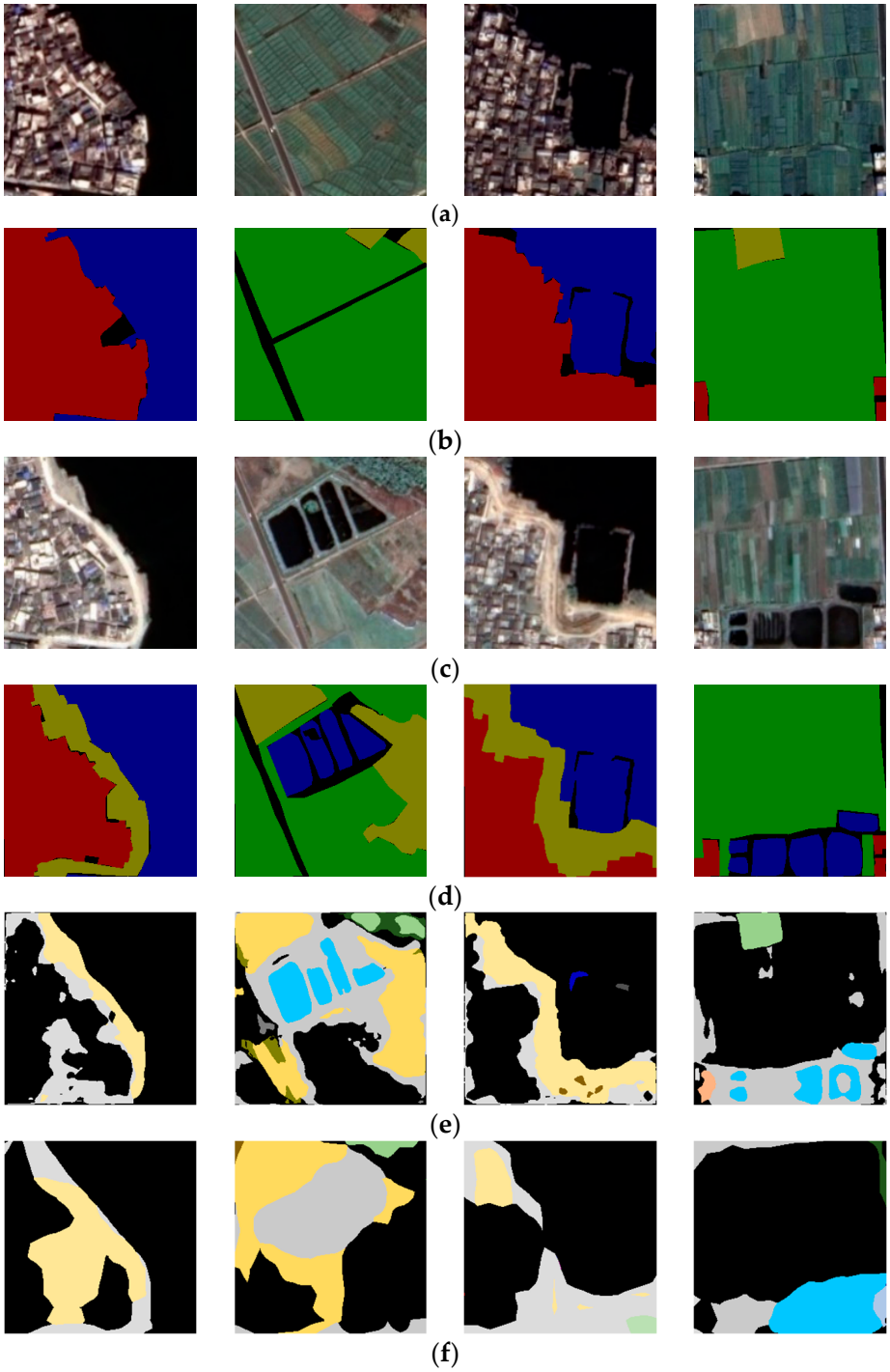

Figure 10. *Cont.*

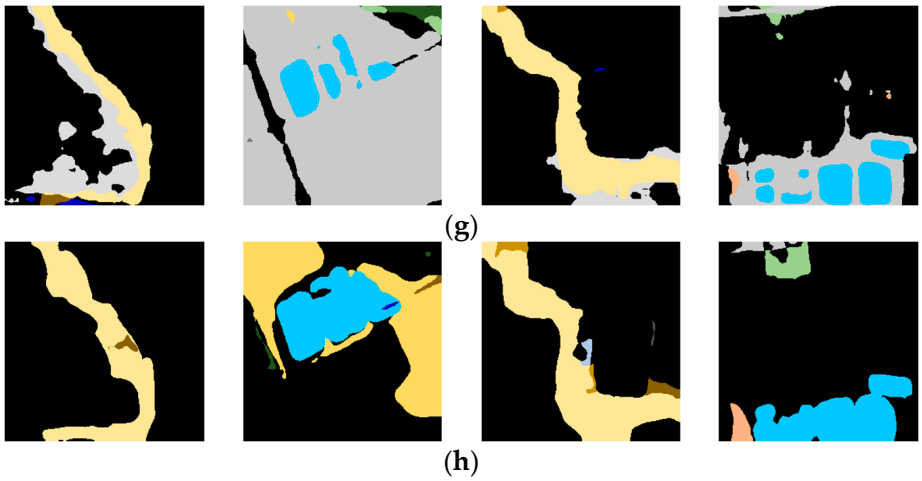

Figure 10. Experiments of different methods for change detection. (**a**) T1 image. (**b**) Ground truth of T1 image. (**c**) T2 image. (**d**) Ground truth of T2 image. (**e**) Siam-U-Net. (**f**) Siam-PSPNet. (**g**) Siam-DeepLabv3+. (**h**) Our method.

Table 4. Comparison of different change detection methods. (%).

Method	*Accuracy*
Siam-UNet	81.2
Siam-PSPNet	75.1
Siam-DeepLabv3+	84.5
Our method	86.0

6. Conclusions

According to the requirement of water management, we use the high resolution remote sensing images of Lake Erhai in Yunnan Province in China collected by GF-2 to make the dataset with a fine-grained semantic label and an image-pair label. There are five classes and 21 semantic changes in our dataset, which is more fine-grained than the public dataset HRSCD, and the data-making process can be applied to any application. Aiming at the variations of land features in high resolution remote sensing images and the requirement of the refined management, we propose a pixel-level semantic change detection method to solve the fine-grained semantic change detection for HR remote sensing image pairs. We firstly propose a lightweight semantic segmentation network to carry out the pixel-level semantic classification, then use the parameter-sharing SiameseNet as the architecture of our method to obtain the different classifications of the image pair, and finish with the pixel-level change detection based on semantic comparison of the image pair. LightNet consists of four long–short branches and obtains feature information at different scales. The features in each branch are transmitted, fused, and enhanced via channel information enhancement layer or spatial information enhancement layers, and the four upsampling feature maps are finally concatenated to form the input of the Softmax. Our method solves the intra-class inconsistency and inter-class similarity, so it not only achieves end-to-end change detection in remote sensing images, but also helps management find the specific change without any manual semantic comparison. The experiments demonstrate the superiority of our method and the accuracy of LightNet, and the pixel-level semantic change detection methods are up to 89% and 86%, respectively.

Author Contributions: Conceptualization, L.Z. and R.Y.; Methodology, L.Z. and M.X.; Validation, Y.X. and G.W.; Resources, R.S. and G.W.; Data Curation, M.X. and R.S.; Writing—Original Draft Preparation, L.Z. and R.Y.; Writing-Review and Editing, L.Z. and Y.X.; Supervision, L.Z.; Funding Acquisition, L.Z. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2023YFC3006501), and Guangdong Water Technology Innovation Project (grant number 2021-07), the Natural Science Foundation of Jiangsu Province (No. BK20201311), and the National Natural Science Foundation of China (No. 62073120, 42075191, 91847301, 92047203, 2009080).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wei, H.; Jinliang, H.; Lihui, W.; Yanxia, H.; Pengpeng, H. Remote sensing image change detection based on change vector analysis of PCA component. *Remote Sens. Nat. Resour.* **2016**, *28*, 22–27.
2. Brunner, D.; Lemoine, G.; Bruzzone, L. Earthquake damage assessment of buildings using VHR optical and SAR imagery. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2403–2420. [CrossRef]
3. Luo, H.; Liu, C.; Wu, C.; Guo, X. Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery. *Remote Sens.* **2018**, *10*, 980. [CrossRef]
4. Coppin, P.; Jonckheere, I.; Nackaerts, K.; Muys, B.; Lambin, E. Review ArticleDigital change detection methods in ecosystem monitoring: A review. *Int. J. Remote Sens.* **2004**, *25*, 1565–1596. [CrossRef]
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
6. Daudt, R.C.; Le Saux, B.; Boulch, A.; Gousseau, Y. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Underst.* **2019**, *187*, 102783. [CrossRef]
7. He, Y.; Zhang, H.; Ning, X.; Zhang, R.; Chang, D.; Hao, M. Spatial-temporal semantic perception network for remote sensing image semantic change detection. *Remote Sens.* **2023**, *15*, 4095. [CrossRef]
8. Muchoney, D.M.; Haack, B.N. Change detection for monitoring forest defoliation. *Photogramm. Eng. Remote Sens.* **1994**, *60*, 1243–1252.
9. Mondini, A.C.; Guzzetti, F.; Reichenbach, P.; Rossi, M.; Cardinali, M.; Ardizzone, F. Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images. *Remote Sens. Environ.* **2011**, *115*, 1743–1757. [CrossRef]
10. Schoppmann, M.W.; Tyler, W.A. Chernobyl revisited: Monitoring change with change vector analysis. *Geocarto Int.* **1996**, *11*, 13–27. [CrossRef]
11. Du, P.; Wang, X.; Chen, D.; Liu, S.; Lin, C.; Meng, Y. An improved change detection approach using tri-temporal logic-verified change vector analysis. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 278–293. [CrossRef]
12. Baronti, S.; Carla, R.; Sigismondi, S.; Alparone, L. Principal component analysis for change detection on polarimetric multitemporal SAR data. In Proceedings of the 1994 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'94), Pasadena, CA, USA, 8–12 August 1992; pp. 2152–2154.
13. Celik, T. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
14. Zhang, M.; Shi, W. A feature difference convolutional neural network-based change detection method. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7232–7246. [CrossRef]
15. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
16. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
17. Ke, Q.; Zhang, P. MCCRNet: A multi-level change contextual refinement network for remote sensing image change detection. *ISPRS Int. J. Geo Inf.* **2021**, *10*, 591. [CrossRef]
18. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7296–7307. [CrossRef]
19. Pang, S.; Li, X.; Chen, J.; Zuo, Z.; Hu, X. Prior Semantic Information Guided Change Detection Method for Bi-temporal High-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1655. [CrossRef]
20. Xiang, S.; Wang, M.; Jiang, X.; Xie, G.; Zhang, Z.; Tang, P. Dual-task semantic change detection for remote sensing images using the generative change field module. *Remote Sens.* **2021**, *13*, 3336. [CrossRef]
21. Xia, H.; Tian, Y.; Zhang, L.; Li, S. A deep siamese postclassification fusion network for semantic change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622716. [CrossRef]
22. Ding, L.; Guo, H.; Liu, S.; Mou, L.; Zhang, J.; Bruzzone, L. Bi-Temporal semantic reasoning for the semantic change detection in HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620014. [CrossRef]
23. Zheng, Z.; Zhong, Y.; Tian, S.; Ma, A.; Zhang, L. ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 228–239. [CrossRef]
24. Yang, K.; Xia, G.S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; Zhang, L. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5609818. [CrossRef]

25. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; He, P. SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102465. [CrossRef]
26. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8007805. [CrossRef]
27. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607514. [CrossRef]
28. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
29. Ying-ming, H.; Feng, Z. Fast algorithm for two-dimensional otsu adaptive threshold algorithm. *J. Image Graph.* **2005**, *10*, 484–488.
30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
31. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.
32. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
33. Zhang, L.; Fan, Y.; Yan, R.; Shao, Y.; Wang, G.; Wu, J. Fine-grained tidal flat waterbody extraction method (FYOLOv3) for High-Resolution remote sensing images. *Remote Sens.* **2021**, *13*, 2594. [CrossRef]
34. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
35. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
36. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
37. Zhang, L.; Liao, Y.; Wang, G.; Chen, J.; Wang, H. A Multi-scale contextual information enhancement network for crack segmentation. *Appl. Sci.* **2022**, *12*, 11135. [CrossRef]
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE international Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
39. Zhang, L.; Wu, J.; Fan, Y.; Gao, H.; Shao, Y. An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN. *Sensors* **2020**, *20*, 1465. [CrossRef] [PubMed]
40. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
41. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
42. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
43. Lippitt, C.D.; Zhang, S. The impact of small unmanned airborne platforms on passive optical remote sensing: A conceptual perspective. *Int. J. Remote Sens.* **2018**, *39*, 4852–4868. [CrossRef]
44. Zhang, S.; Lippitt, C.D.; Bogus, S.M.; Loerch, A.C.; Sturm, J.O. The accuracy of aerial triangulation products automatically generated from hyper-spatial resolution digital aerial photography. *Remote Sens. Lett.* **2016**, *7*, 160–169. [CrossRef]